A Multiple Instance Learning and Relevance Feedback Framework for Retrieving Abnormal Incidents in Surveillance Videos

Chengcui Zhang, Wei-Bang Chen, Xin Chen, Lin Yang, and John Johnstone
Department of Computer and Information Sciences, The University of Alabama at Birmingham, AL, USA
Email: {zhang, wbc0522, chenxin, galabing, jj}@cis.uab.edu

***Abstract*—This paper incorporates coupled hidden Markov models (CHMM) with relevance feedback (RF) and multiple-instance learning (MIL) for retrieving various abnormal events in surveillance videos. CHMM is suitable for modeling not only the object's behavior itself but also the interactions between objects. In addition, to address the challenges posed by the "semantic gap" between high level human concepts and the machine-readable low level visual features, we introduce relevance feedback (RF) to bridge the semantic gap by progressively collecting feedback from the user, which allows the machine to discover the semantic meanings of an event by exploring the patterns behind low-level features. The adopted multiple-instance learning algorithm enables the proposed framework to provide a user-friendly video retrieval platform with the use of query-by-example (QBE) interface. The experimental results show the effectiveness of the proposed framework in detecting "chasing", "fighting", and "robbery" events by demonstrating the increase of retrieval accuracy through iterations and comparing with other methods. By tightly integrating these key components in a learning system, we ease the surveillance video retrieval problem.**

***Index Terms*—video retrieval, relevance feedback, coupled hidden Markov models, multiple instance learning, surveillance videos, spatio-temporal modeling**

I. Introduction

The surveillance camera system has been widely used to ensure site security, which not only deters criminals but also provides solid evidence for settling disputes or investigating crimes. The interpretation of surveillance videos heavily relies on human, which requires dedicated personnel spending hours or days, to sequentially screen and review the entire video sequence. The event extraction from surveillance videos becomes even more difficult when videos are recorded from multi-cameras since images from different cameras may be displayed concurrently in one screen or may be interleaved in a round-robin fashion. Therefore, it is essential to automate the process of surveillance so that the time and labor cost can be greatly reduced. In recent years, the video recording technique has been transitioning from analog to digital, which enables and motivates researchers to develop more intelligent surveillance systems. The intellectual surveillance system collects huge amount of surveillance videos via security cameras and stores them in the database, which demands effective and efficient surveillance video retrieval to make the information extraction (e.g., event extraction) from surveillance videos feasible.

A lot of works exist in detecting and recognizing events in videos. Many studies in this area are solely based on the generic visual properties of video frames [1]. However, these works do not adopt the spatio-temporal information. Many other works adopt object trajectories as the basis for analysis. However, these approaches mainly focus on the decomposition and approximation of motion trajectories, where semantic meaning about object motions is not reflected. Aside from that, none of these approaches involve any learning process in them [2, 3]. Stochastic methods were also exploited in learning and recognizing video events [4-8]. Our proposed framework adapts one of the stochastic methods, i.e., CHMM (Coupled Hidden Markov Model), for detecting abnormal human interactions in the indoor surveillance videos. In addition, Self Organization Map (SOM) has also been used in some works for event detection from videos [9]. Our proposed learning framework is different from [9] in that our input is time series sequences with temporal constraints.

A typical video database query in such applications would be to retrieve abnormal events with high-level semantic meanings, such as robbery and fighting scenes. In order to bridge the semantic gap, Hu et. al [2] proposes a framework that learns the incident models by clustering trajectories hierarchically with the use of spatial and temporal information [2]. Very few other frameworks borrow the concept of relevance feedback (RF) from the content-based image retrieval (CBIR) to learn the incident models in a progressive fashion, including our previous work [10] and the work proposed by Meessen et al. [11]. As a supervised learning technique, relevance feedback incorporates the subjective perceptions from users with the learning process, which significantly increases the retrieval accuracy.

Though CHMM provides a better representation of human interactions [5], it still requires proper features to approximate event models. It is worth mentioning that we not only adopt conventional motion information to model the macro interactions between objects, but also incorporate local features such as optical flow to capture the micro interactions between objects. More specifically,

macro interactions profile an event at a coarse granularity, while the micro interactions detail the event at a finer granularity. By adopting both macro and micro interactions in modeling events, the proposed framework is capable of differentiating events such as "meet-handshake-split", "meet-fight-runaway", and "meet-robbery-runaway". All the above events involve "two people get together and split"; however, their meanings and actual forms of presence in the video are quite different.

In order to extract proper input features from the surveillance video, it is necessary to isolate objects from the video as segments and track their positions as object trajectories. Many segmentation algorithms have been proposed for tracking objects in the video [12-15], which go beyond the scope of this paper. In our previous work [12], we proposed an object segmentation and tracking algorithm for surveillance videos, from which object-level information such as the bounding boxes and the centroids can be obtained and stored in the database for future queries. This object segmentation method is adopted in this paper for video pre-processing, followed by a manual data cleaning step to remove the noise and trajectory distortions incorrectly introduced by inaccurate object segmentation. Once objects are segmented and their trajectories have been generated, we subsequently divide the entire video into small segments which should ideally be small enough for efficient information retrieval and indexing and big enough to enable effective object analysis. In general, for indexing purposes, video can be segmented into shots according to sharp scene changes. However, surveillance videos lack this property since they are composed of monotonously running frames. Therefore, we adopt Common Appearance Interval (CAI) in our framework as a video segmentation concept which endows a segment of video with semantic meaning in terms of temporality and spatial relations [16]. In brief, a new video segment is generated whenever a new object enters or leaves the surveillance area covered by the camera. We exemplify the concept of CAI in Fig. 1.

After organizing the entire video sequence into CAIs, we study pairs of object trajectories, which will be referred to as Trajectory Pair (TP). Further, in order to incorporate CHMM into the proposed framework, a sliding window which sequentially moves along a TP is used to extract fixed length of shorter trajectories from the TP, which will be referred to as Sequence Pair (SP) in this paper. By analyzing each SP, we provide a way to model events that involve multiple people interactions. This is the key point of this paper to model events that heavily involve object interactions, as opposed to the existing study on event retrieval by analyzing the behavior of individual objects. The latter has been well-studied in the literature [17-19].

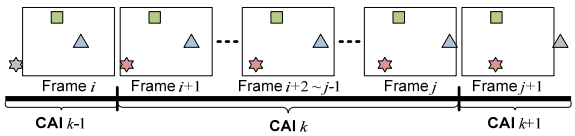

Figure 1. Common appearance interval (CAI).

In order to provide a user-friendly semantic video retrieval system, the proposed framework offers query-by-example (QBE) features at the CAI level. This system allows a user to provide a CAI as a query example which contains an instance of the event of interest. The system returns CAIs that contain at least one similar instance of the event of interest as appearing in the query CAI. The challenge is that a CAI may contain events other than the query event, while the user's feedback is provided for the entire CAI instead of for individual events. For example, a third party may pass by a scene of two people fighting, but the event of his walking through a fighting scene is not what the user is interested. However, by incorporating the relevance feedback from the user, the learning system is able to guess the actual event of interest to some extent. To address this challenge, we incorporate Multiple-Instance Learning (MIL) with relevance feedback in our framework.

The multiple-instance learning (MIL) problem is a special kind of supervised machine learning problem. In standard supervised machine learning scenario, training data are individually labeled in order to train a classifier for predicting the class of unseen objects. However, in multiple-instance learning scenario, the class labels of individual objects in the training set are not available. Instead, users provide class label to a set of objects called a bag, while an individual object in a bag is called an instance. In other words, a bag with a negative label indicates that there is no related instance in the bag, while a bag with a positive label implies that there is at least one related instance in the bag. The multiple-instance learning algorithm utilizes the class label from bags for predicting the class label for unseen bags as well as instances. The incorporation of relevance feedback and multiple-instance learning perfectly matches the surveillance video retrieval scenario [10, 11] since the former reduces the semantic gap by incorporating the user's high-level perception and gathering training samples in a progressive manner, while the latter guesses the event of interest by analyzing the collected training samples based on incomplete training label information. It should be pointed out that the proposed framework is different from [11] in that our framework not only incorporates RF and MIL into the retrieval, but also tightly integrates them with CHMM and takes advantage of CHMM to model various kinds of human interactions. Another difference is that in this study, we explore features that can capture 'micro interactions.' 'Micro interactions' can be very important indicators for differentiating various types of events. To our best knowledge, our work is among the first few to tackle the surveillance video retrieval problem this way in one single integrated system.

In summary, the proposed framework contributes to a "human-centered" surveillance video retrieval system which integrates multimedia processing, spatiotemporal modeling, multimedia data mining, and information retrieval techniques. This framework analyzes macro and micro interactions between pairs of objects in an example CAI and retrieves CAIs with events of interest according

to individual users' query interest. Currently, the proposed framework can recognize several kinds of abnormal events, including 'fighting', 'chasing', and various types of 'robbery' events. It is worth mentioning that the proposed framework can be easily tailored to the recognition of other abnormal interactions.

In the remainder of the paper, Section II describes the design detail of the proposed framework. Section III presents the experimental results. Section IV concludes the paper.

II. The Proposed Framework

Fig. 2 illustrates the high-level architecture of the proposed framework. In this section, we will introduce the design detail of each component, including video preprocessing, event modeling, and event learning and retrieval.

A. Video Preprocessing

As we stated in Section 1, a surveillance video sequence is composed of monotonously running frames, and thus, no hard boundary or significant change of background can be used for segmenting the video into shots. However, we observe that an event of interest, in general, involves at least one human object. Therefore, based on this observation, we incorporate the concept called Common Appearance Interval (CAI) into our framework [16]. A CAI is defined as an interval where a certain set of objects appear in a sequence of video frames together. In this way, surveillance videos are indexed based on the common appearance/disappearance of video objects.

Fig. 1 illustrates the video segmentation schema used in the proposed framework. As shown in the example, Frame i contains a 'square' object and a 'triangle' object. A 'star' object enters Frame i+1. The entering of this new star object terminates the k-1th CAI and starts a new CAI k. Similarly, as the triangle object moves towards the edge of the Frame j and disappears in Frame j+1, its disappearance simultaneously signifies the ending of CAI k and the starting of a new CAI k+1. Therefore, a new CAI is generated whenever a new object enters the scene or an existing object leaves the scene. To this end, surveillance video are efficiently indexed and stored in the database for future access.

B. Camera Calibration for 2D to 3D Back Projection

One disadvantage of 2D trajectories is that their metric is not invariant to perspective distortion. For example, the speed of objects appears to be lower in the image as they move further away from the camera. The ability to reconstruct 3D object trajectories from 2D image sequences not only improves the accuracy of object trajectory extraction, but also reduces the dependency of the system on a specific environmental setting.

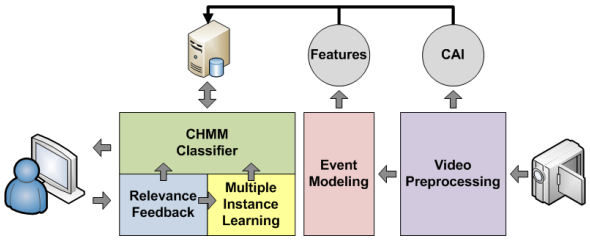

Figure 2. The architecture of the proposed framework.

We approach this problem by using a structure from motion (SfM) framework. The camera parameters are calibrated using an off-line process, where at least 6 points (11 points in our case) on an image with known 3D coordinates (not on the same plane) are manually selected to calibrate the camera parameters using the DLT algorithm [20]. The camera parameters are used during on-line tracking to reconstruct the 3D coordinates of objects in the image.

3D reconstruction of image points from a single view is an under-constrained problem, even with known camera parameters, because the 3D coordinate of an image point can lie anywhere on the 3D ray emitted from the camera center through the image point. In order to gain more constraints, we adopt a heuristic from [21] to estimate the ground projection of an image point and solve its 3D coordinate in two steps.

In the first step, we find the ground projection of an image point by dropping a plumb line down to the bounding box of the object and treating the intersection as its ground projection. By assuming a zero Z-coordinate on the ground, we can solve the 3D coordinate of the ground projection by intersecting the ray with Z=0 plane. Since a 3D point shares the same X/Y-coordinate as its ground projection, we can solve the 3D coordinate of the off-ground point once its ground projection is solved. Detailed equations for the above procedure are given below.

Let $P=(P_1\ P_2\ P_3\ P_4)$ be the 3×4 camera matrix given by the DLT algorithm where P_1, P_2, P_3 and P_4 are the four columns, $\boldsymbol{x}=(x\ y\ 1)^{\mathrm{T}}$ be an image point to reconstruct, and $\boldsymbol{x}'=(x'\ y'\ 1)^{\mathrm{T}}$ be its ground projection on the image (points are in Homogenous coordinates). In the first step, we solve the 3D coordinate $\boldsymbol{X}'=(X'\ Y'\ Z'\ 1)^{\mathrm{T}}$ for $\boldsymbol{x}'$:

$$\begin{cases} P\boldsymbol{X}' = k\boldsymbol{x}' \\ Z' = 0 \end{cases} \tag{1}$$

where k is the scale factor in Homogenous space. Then X' and Y' are given by:

$$\begin{pmatrix} X' & Y' & 1 \end{pmatrix}^{\mathrm{T}} = k\begin{pmatrix} P_1 & P_2 & Z'P_3 & P_4 \end{pmatrix}^{-1}\boldsymbol{x}' \tag{2}$$

Since $\boldsymbol{X}$ and $\boldsymbol{X}'$ share the same X/Y-coordinate, $\boldsymbol{X}=(X\ Y\ Z)$ is given by:

$$\begin{cases} P\boldsymbol{X} = k\boldsymbol{x} \\ X = X_0 \end{cases} \tag{3}$$

where X_0 is the solution for $\boldsymbol{X}'$. As shown in Equation (4), Equation (3) can be solved in the same way as Equation (1):

$$\begin{pmatrix} Y & Z & 1 \end{pmatrix}^{\mathrm{T}} = k\begin{pmatrix} P_1 & P_2 & X_0P_3 & P_4 \end{pmatrix}^{-1}\boldsymbol{x} \tag{4}$$

Reconstructing 3D trajectories from 2D image sequences effectively corrects the perspective distortion, which allows the extraction of more accurate object trajectories for the subsequent event modeling and retrieval steps.

C. Event Modeling

As we mentioned earlier, the goal of this research is to detect various abnormal human interactions which often involve the behavior of at least two persons. In order to differentiate events which have similar macro interactions but differ from each other in terms of micro interactions, we have to choose features properly to model various interaction events.

Various motion properties based on object trajectories can be extracted for modeling macro interactions. In this study, we choose the following three features for describing macro interactions, including (1) relative distance (*dist*) that describes the distance between two objects, (2) degree of alignment (θ) which represents the angle difference between the motion vectors of two objects, and (3) change of velocities (*vdiff*) which is the velocity change of two objects between two consecutive sampling frames. In addition, we use (4) motion energy (*me*), which is the magnitude of motion change of each object, to describe micro interactions.

The first three features can be easily obtained from object trajectories. This relative distance (*dist*) feature is defined as the Euclidian distance between the centroids of two objects. In order to calculate the degree of alignment (θ), we first need to find the motion vector for each object. The motion vector is actually the coordinate difference of object centroids in adjacent frames. The degree of alignment is thus the signed angle between two motion vectors. The change of velocities (*vdiff*) can be obtained by calculating the displacement d of an object in consecutive frames. The velocity is defined as d/t, where t is the time interval between two consecutive frames.

The last feature motion energy (*me*) uses the Optical Flow to measure the magnitude of motion change of each object, which enables us to model the micro interactions between two objects. The Optical Flow calculates the velocity and the direction of the pixel motions within the bounding box areas of that object. The basic idea of Optical Flow is to find the difference between one pixel in the current frame and the corresponding point it moves to in the next frame, which raises the problem of correlating corresponding points/pixels between two consecutive frames.

The Optical Flow technique is based on the assumption that images are made of patches whose intensities change smoothly. Therefore, for a pixel in the image, its surrounding pixels have similar intensities. Therefore, we can obtain Optical Flow of point F by minimizing the intensity differences within the bounding box areas from two consecutive frames. The motion energy feature (*me*) of an object in its bounding box at time t is defined as the mean flow norm:

$$me = \frac{\sum \|F\|}{N(t)}$$

where, N is the total number of pixels in the bounding box at time t.

In this study, an event can be considered as a series of interactions between objects in time. Therefore, to differentiate between different events, we should consider not only the interaction between objects at a certain time, but also their spatio-temporal relations or how the interaction evolves over time. In this paper, we quantify an interaction with four features, i.e., relative distance (*dist*), degree of alignment (θ), change of velocities (*vdiff*), and motion energy (*me*), and describe the temporal relations with a collection of feature vectors in a sliding window. For example: In a fighting event, two persons must keep their relative distance short for a period of time in order for the model to learn and identify fighting events. In addition, the motion energy must remain high due to the fighting actions. However, in a robbery event, a robber first approaches a victim, which is reflected by the decreasing relative distance between the two subjects, and then, the robber escapes from the scene, which is signified by the increasing relative distance. The high motion energy only occurs at the moment of robbing, and remains low during the rest of the time. This is especially true in an indoor surveillance environment where the area monitored by a single camera is relatively small such that a robbery event lasts for only a few seconds. Our experimental results demonstrate that the spatio-temporal features aforementioned are reasonably effective in differentiating between different events.

Thus, any object in the video at time t can be characterized with a four-feature vector v, and consequently, a SP (or TP) which consists of two objects a and b can be represented as a series of such vectors: $SP_a = [va_1, va_2, \ldots, va_n]$ and $SP_b = [vb_1, vb_2, \ldots, vb_n]$.

D. Event Learning and Retrieval

1) Coupled Hidden Markov Models (CHMM): Hidden Markov Model (HMM) is a stochastic process that automatically performs dynamic time warping for time-series data. The HMM considers a system that consists of a finite set of states, each of which is connected by transitions with associated probabilities. The transitions between the states are governed by the transition probabilities, and thus, convey a clear Bayesian semantics. Fig. 3 exemplifies a two-state HMM with an example of a rolled out observation sequence modeled by the HMM.

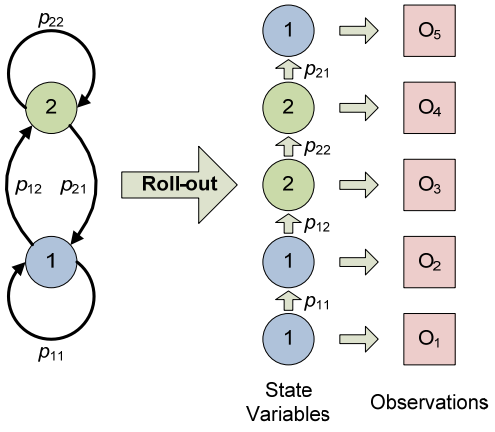

Figure 3. A two-state HMM with rolled-out observations.

where p_{ij} represents the transition probability from state i to state j.

In this study, our goal is not only modeling an object itself, but also the interactions between objects. Thus, traditional hidden Markov models are ill-suited for this problem. Instead of using single Markov chain, we introduce Coupled Hidden Markov Models (CHMMs) which allow for modeling a stochastic process with more than one state during a single time interval. We illustrate the tree structure of a CHMM rolled-out in time in Fig. 4.

As shown in Fig. 4, a CHMM is essentially a series of parallel HMM chains coupled through adding transition probabilities cross time and cross chain in order to model the interactions between different entities. In this paper, we adopt a two-chain CHMM to model the interaction between two persons in the surveillance video. The posterior probability of a state sequence through a two-chain CHMM is defined as:

$$P(Q \mid O) = \frac{P_{q_1} p_{q_1}(O_1) P_{q_1'} p_{q_1'}(O_1')}{P(O)} \times \prod_{t=2}^{T} P_{q_t \mid q_{t-1}} P_{q_t' \mid q_{t-1}'} P_{q_t \mid q_{t-1}'} P_{q_t' \mid q_{t-1}} P_{q_t}(O_t) P_{q_t'}(O_t')$$

where q_t, q_t' denote states and o_t, o_t' denote observations in objects A and B, respectively; t represents the t^{th} state over time and T is the length of the observation and state variable sequence. The two-chain CHMM problem can be solved with the use of N-head dynamic programming in (MN^4), and further, in $O(4MN^2)$ by relaxing the assumption that every transition must be visited [22].

In this paper, we model the behavior of a person through one-chain in CHMM, while the interactions between two persons are reflected in the cross transitions between two chains. Therefore, both the individual behaviors and the interactions between two persons can be modeled in a single system.

2) Extracting Sequence Pairs: As a matter of fact, the surveillance video can be viewed as a kind of time-series data with abundant spatiotemporal information. Analyzing data of this kind, we shall not only focus on each individual data point separately but also take their continuity into account, which motivates us to use a sliding window for procuring consecutive yet overlapped data sequences in the proposed framework. Conceptually, the sliding window is similar to the use of microfilm reader browsing data recorded on the microfilm. In the proposed framework, we use a window of a fixed size which moves along the surveillance video in a sequential fashion, and collects segments of object trajectories from the video for subsequent learning and retrieving. The use of sliding window not only allows us to procure consecutive data but also enables us to connect to CHMM model which accepts only finite states.

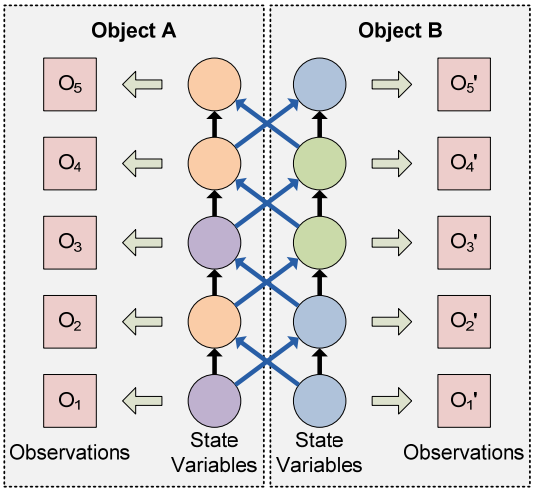

Figure 4. A two-chain CHMM roll-out in time.

We illustrated the concept of sliding window in Fig. 5. This example demonstrates a sliding window of size 5 moving along a time-series data with step size 2, i.e., shifting the window to the right by 2 data points. In addition, 5 data points are collected in each move.

In this study, the surveillance videos we used for experiments are taken at 25 fps, and we collect data points with a sampling rate of 5 frames, i.e., each frame represents 40ms and each sampling data point represents 200ms. The selection of this particular sampling rate is actually a trade-off between data pre-processing efficiency and data precision. We empirically set the sampling rate to 5 frames since a relatively significant object movement, if any, can be observed and detected within 5 frames without losing too many trajectory details.

In the proposed framework, we applied a sliding window of size 10 with step size 1 on the sampled data points, which covers 50 frames corresponding to 2 seconds. The window size is experimentally determined. According to our observations, a too small window may not have enough coverage of the interactions in an event, while a too big window is not appropriate for the detection of shorter events which contain much less frames. In addition, as stated in Section I, the object trajectories obtained from the sliding window are referred to as sequence pairs (SP). This processing is applied to each trajectory pair in each CAI.

3) The Initial Query: In the initial query, the user specifies a CAI with an event of interest as the query target. The goal is to retrieve those CAIs that contain similar events. At this point, the system has no knowledge at all regarding the user's event of interest or the feedback information from the user. In order to provide an initial set of CAIs for the user to provide relevance feedback, the proposed system initiates the learning and retrieval process with a heuristic search.

In the heuristic search, the proposed system builds a CHMM model for each TP (Trajectory Pair) in the query example CAI and uses all the SPs in that TP as training data. Subsequently, the set of trained CHMM models are used for testing against all the SPs in the database.

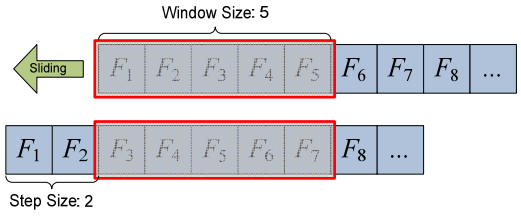

Figure 5. The concept of sliding window.

Assume there are m models. For each model i, we rank CAIs in the database according to their corresponding model posterior probabilities, and therefore, each CAI is assigned a rank r_i associated with the model i. The overall rank of a CAI is calculated as the sum of all its r_i (i=1…m). The proposed system then returns the top 20 CAIs according to the overall ranking. The reason that we calculate similarity scores in this way is that the use of CHMM involves random initialization in the training stage, which makes the posterior probabilities returned from different models not directly comparable in the testing stage. If we fix the initialization for all models, it may increase the chance of a biased initial setting that favors certain models but produces poor local optima for the others.

After the initial query, a certain number of SPs are presented to the user in the form of CAIs. In our experiment, the top 20 unique CAIs are returned for the user's feedback. The user can now review the retrieved CAIs by clicking the 'play' button. If a CAI contains the query event, the user marks the CAI as 'relevant' by checking the check box below that CAI. With the user's feedback at hand, a set of training samples can be collected.

4) Open Two-Level Multiple Instance Learning: In multiple instance learning, unlike the case for traditional supervised learning, the label of an individual object is unknown. Instead, only the label of a set of objects is available. An individual object is called an instance and a set of instances with an associated label is called a bag. Specifically, in video retrieval with RF, there are usually only two kinds of labels, namely Positive and Negative. A bag is labeled Positive if the bag has one or more than one positive instances and is labeled Negative if and only if all of its instances are negative.

The Multiple Instance Learning problem requires learning a function mapping from an instance to a label (either Positive or Negative) with the best possible approximation to the unknown real mapping function.

In our proposed framework, the label space is transformed from a discrete space {1 (Positive), 0 (Negative)} to a continuous space [0, 1] since the proposed framework is a retrieval framework rather than a classification framework. Therefore, an instance in the database is more or less relevant to the query instance, and its similarity to the query instance is represented as a value in a continuous space [0, 1]. In the proposed Multiple Instance Learning framework, the label of a bag actually indicates the extent to which that bag is Positive – instead of either one hundred percent Positive or Negative. The label '1' means the bag is one hundred percent Positive, while '0' indicates that the bag is zero percent Positive. The same applied to the label of an instance. The goal of learning subsequent to this transformation is to generate a mapping function f from the training examples to predict the extent to which an instance is positive. In addition, the extent to which a bag is positive is determined by the maximum posterior probability that its instances are positive. In this study, instead of using a fixed cutoff value to separate positive instances from the negative ones, the top 99.5% highest scoring SPs (from positive CAIs) in terms of their posterior probabilities are selected as positive instances for subsequent model training. The reason for selecting positive instances with this criterion is due to the fact that the proposed learning mechanism, when coupled with relevance feedback, is a so-called small supervised learning approach where the training data set is very small. In such a situation, the trained model, especially during the first one or two iterations of RF, is often inaccurate. Consequently, the posterior probabilities of most SPs in the user selected CAIs (positive CAIs) are way below 0.01 in such cases. Therefore, it is not proper to use a hard cutoff value such as 0.5 to differentiate positive instances from negative ones. In addition, selecting the top 99.5% highest scoring SPs can eliminate extreme cases according to the statistical theory, which may help in reducing possible noise from irrelevant SPs.

The collection of training samples collected in our CAI-based video retrieval actually poses a two-level multiple-instance learning problem. At the first level, each retrieved CAI can be considered as a bag, and all the TPs in a CAI can be considered as its instances. At the second level, each TP becomes a bag, and all the SPs in that TP become instances. From user's feedback, we can obtain the class label (bag label) for each retrieved CAI. In MIL scenario, if a CAI is marked as "relevant", it indicates that at least one TP in that CAI contains the event of interest. Further, if a TP contains an event of interest, at least one SP in that TP is similar to the event of interest. This implies that a SP with the maximum posterior probability is likely to be the event of interest.

To transform this two-level MIL to traditional supervised learning, the system needs to determine instance labels from bag labels at both levels. At the lowest level, for a given positive CAI, we first identify its SP that results in the highest posterior probability according to the current CHMM model and given it a positive label which is then used as the label for the TP that contains the SP. Ideally, only the highest-scored SPs from all Positive CAIs should be included in the training set. However, since the number of training SPs that can be collected is usually very small (given that only the top 20 CAIs are returned), we procure all the SPs whose posterior probability is above a certain threshold (e.g., top 99.5% highest scoring SPs) in the highest-scored TP in each CAI as training samples and feed them into the learning algorithm, which learns the best parameters for the CHMM.

In the subsequent retrieval-feedback iterations, these parameters are further refined with new training samples collected from users' feedbacks. In this iterative process, the user's query interest is obtained as user feedbacks and transferred to the learning algorithm, and the refined results are returned to the user for the subsequent run of the retrieval-feedback and multiple-instance learning. It is shown in our experiment that, with this interactive learning technique, the retrieval results can be improved iteratively.

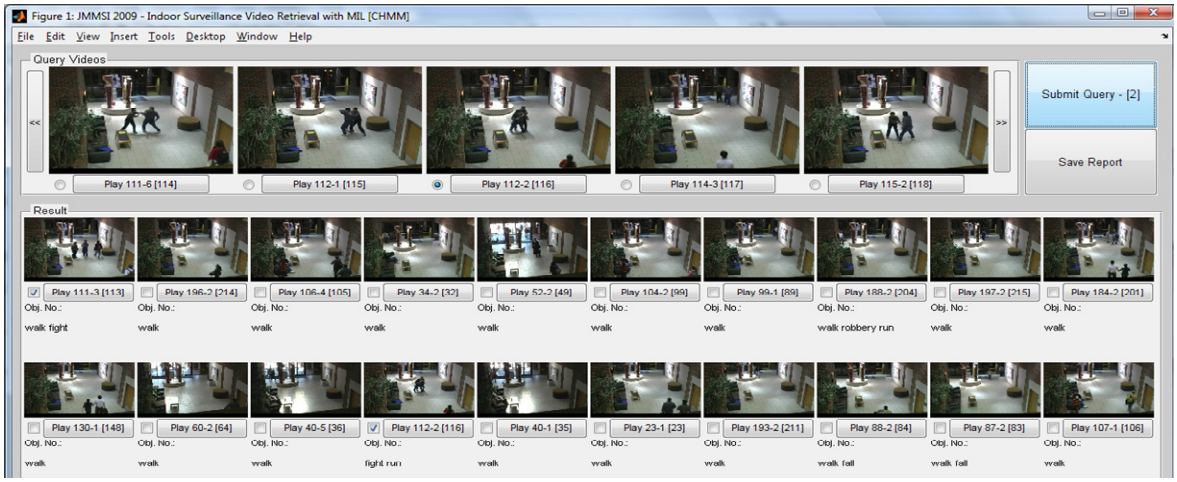

Figure 6. The system query interface.

It is worth noting that here still exists a major difference between Multiple Instance Learning and traditional supervised learning despite the transformation from Multiple Instance Learning to the traditional supervised learning. The training set is static and usually does not change during the learning procedure in traditional supervised learning. However, in the transformed version of Multiple Instance Learning with RF, the training set may change during the learning procedure. The reason for this is that the instance with the maximum label in each bag may change with the update of the mapping function f during the learning procedure. The training set constructed along with the aforementioned transformation may also change during the learning procedure. The fundamental learning method remains the same despite such dynamic.

III. Experimental Results

A. Dataset Description

The proposed system has been tested on real surveillance data which is procured at the lobby of the Campbell Hall at the University of Alabama at Birmingham (UAB). The original surveillance video corpus is taken at a frame rate of 25 fps, and its total length is about 360,000 frames (4 hours). In the entire video corpus, there are only 63,569 frames (42 minutes) with the existence of objects. Those frames without object are automatically removed from the corpus after object segmentation and trajectory detection. This pre-processing step greatly reduces the search scope, and therefore, reduces time-complexity. This process also effectively prevents possible noise due to the irrelevant video content. After pre-processing, the experimental video corpus consists of 63,569 frames, 677 CAIs, 1,148 TPs, and 9,133 SPs.

A ground truth has been manually established to identify several scenarios involving multiple objects, such as fighting (34 CAIs), chasing (33 CAIs), and robbery (44 CAIs). The robbery events can be further classified into three subtypes, including Type I - "A robbed B and ran away" (18 CAIs), Type II - "A&B robbed C and ran away" (13 CAIs), and Type III - "A&B stop C, robbed C, and ran away" (13 CAIs).

B. User Interface

Fig. 6 demonstrates the interface of the proposed surveillance video retrieval system. On the top of the interface, user can specify the query target by selecting an example CAI. Once a query is submitted, the proposed system returns the top 20 CAIs to the user according to the method presented in Section D.3 (The Initial Query). The user can review the retrieved CAIs by clicking the 'play' button. If a CAI contains the query event, the user marks the CAI as 'relevant' by checking the check box below that CAI. As shown in Fig. 6, 2 CAIs are labeled as relevant to the fighting event. However, at this point, the system has no knowledge of the actual query event, i.e., fighting events.

C. Performance Measurement

Instead of using precision-recall, which is a widely used evaluation methodology in CBIR, we use the measure of accuracy and recall in this research for evaluating performance and comparison purpose.

Accuracy serves as the standard by which to measure the retrieval performance of a CBIR system. The term accuracy is defined the same as precision within a certain scope. In particular, the accuracy rates within different scopes, i.e., the percentage of relevant CAIs within the top 5, 10, 15 and 20 returned CAIs are calculated.

Recall measures the probability that a relevant document is retrieved in a search and is defined as the ratio of number of relevant CAI retrieved and the number of relevant CAI in the top n retrieved results, where n is the number of relevant CAI in the dataset.

D. System Performance

In this study, abnormal human interactions are modeled for indoor surveillance video retrieval. We tested various types of events with the proposed framework. These events include "fighting", "chasing", and "robbery" events. As mentioned earlier in this section, the "robbery events can be further divided into three sub-classes.

In our experiment, five rounds of user relevance feedback are performed - Initial (no feedback), First, Second, Third, and Fourth. During each iteration, the top 20 unique CAIs are returned to the user in the form of video segments.

Usually, the number and type of users could affect the results in most retrieval framework with relevance feedback. However, in our particular case of video event retrieval, such impact is ignorable due to the following reasons: 1) First, the abnormal events in this study can be easily differentiated. A user can easily tell the difference between fighting, chasing, and different types of robbery events as they are well defined. Therefore, the feedback from users is almost always consistent among all the users being tested. 2) Based on our observations from (1), in order to evaluate the system performance in a comprehensive way, we manually examine each CAI and label all the events involved in that CAI as the ground truth in our dataset. In our experiments, given a query CAI and the event of interest it contains, the relevance feedback information is automatically collected from the top few returned CAIs according to the corresponding ground truth. The overall system performance reported for each type of event is thus the averaged result over all query CAIs in that event category.

In our experimental design, we compare the proposed CHMM framework with two existing algorithms, including the Hidden Markov Models (HMM) and the traditional Weighted Relevance Feedback (WRF). CHMM and HMM are both graphical models and have the ability to model temporal relations. As mentioned in Section II D.1, a CHMM is essentially a series of parallel HMM chains coupled through adding transition probabilities cross time and cross chain. By comparing these two models, we can better understand the effectiveness of CHMM in modeling the interaction between objects. It is worth noting that for a HMM, each SP is represented by a series of seven-feature vectors SP = $[1/dist_t, \theta_t, \theta'_t, vdiff_t, vdiff'_t, me_t, me'_t]$ under the constraint that the HMM models each SP as a 7-channel sequence instead of two multi-channel sequences as in CHMM. In addition, in order to see how effective CHMM coupled with RF is in learning event models from videos, we further compare the proposed framework with WRF technique which is commonly used in content-based image retrieval systems. The feature vector used in WRF is the same as that used in HMM for comparison. Another reason we compare our framework with WRF is that WRF is a non-graphical method which is structurally different from CHMM.

In order to ensure a fair comparison, the results of the initial round of retrieval (no feedback) in all the three methods are identical, i.e., we pass the initial retrieval results generated from CHMM to the other two methods. In the subsequent iterations, CHMM treats each SP as two four-feature vectors each of which represents one object in that SP. HMM and WRF both treat each SP as a seven-feature vector. In addition, in the WRF method, feature components are associated with a weight vector. With the user's relevance feedback, the feature vectors of all high-scored relevant SPs from relevant CAIs are collected as training samples. The inverse of the standard deviation of each feature component in the training set is used as the updated weight for this feature component in the next round of retrieval.

We observe that it is essential to normalize the weights since some large weights may introduce bias in calculating relevance scores, and thus, affect the retrieval accuracy. Our first attempt to linearly normalize these weight values between 0 and 1. However, it may introduce weight values of zero, which eliminates the corresponding feature. To tackle on this problem, we tried another method, i.e., the percentage of each weight among the total weight is used as its normalized weight. In our experiment, it is found that the latter outperforms both the linear normalization and no normalization at all.

1) Major Abnormal Events Retrieval: Our dataset contains three types abnormal events, including fighting (34 CAIs), chasing (33 CAIs), and robbery (44 CAIs). To evaluate the performance in retrieving abnormal events, all the CAIs in the same event category are used as query CAIs for that event, one at a time. The experimental results for retrieving a specific event are represented as the averaged accuracy of all the queries associated with that type of events. Figs. 7-9 compare the retrieval accuracies on "chasing", "robbery", and "fighting" events, respectively, among the top 20 CAIs and their recall values over five retrieval-feedback iterations. CHMM denotes the proposed framework; HMM is hidden Markov models; WRF is the weighted relevance feedback method.

For "chasing" events, the accuracy values of CHMM, HMM, and WRF at the 5th iteration are 52.73%, 45.15%, and 45.00%, respectively. The recall values are 37.65%, 35.08%, and 27.27%, respectively. For "robbery" events, the accuracy values of CHMM, HMM, and WRF at the 5th iteration are 66.59%, 60.68%, and 65.00%, respectively. The corresponding recall values are 41.79%, 41.53%, and 38.64%, respectively. For "fighting" events, the accuracy values of CHMM, HMM, and WRF at the 5th iteration are 70.29%, 57.79%, and 60.00%, respectively. The recall values are 46.28%, 41.35%, and 43.86%, respectively.

As can be gleaned from the experimental results, CHMM significantly outperforms HMM and WRF in retrieving chasing and fighting events in terms of both accuracy and recall at the 5^{th} iteration. In retrieving robbery event, CHMM outperforms HMM at the 5^{th} iteration and is slightly better than WRF in terms of accuracy; however, the recall value of CHMM is the highest among all.

Moreover, it can be observed that the accuracy and recall values of CHMM and HMM increase monotonically across iterations, which is not always the case for WRF. This observation indicates the effectiveness of using relevance feedback.

The experimental results also show that the average accuracy values (across all three types of events) of CHMM, HMM, and WRF at the 5^{th} iteration are 63.20%, 54.54%, and 56.67%, respectively. The average accuracy

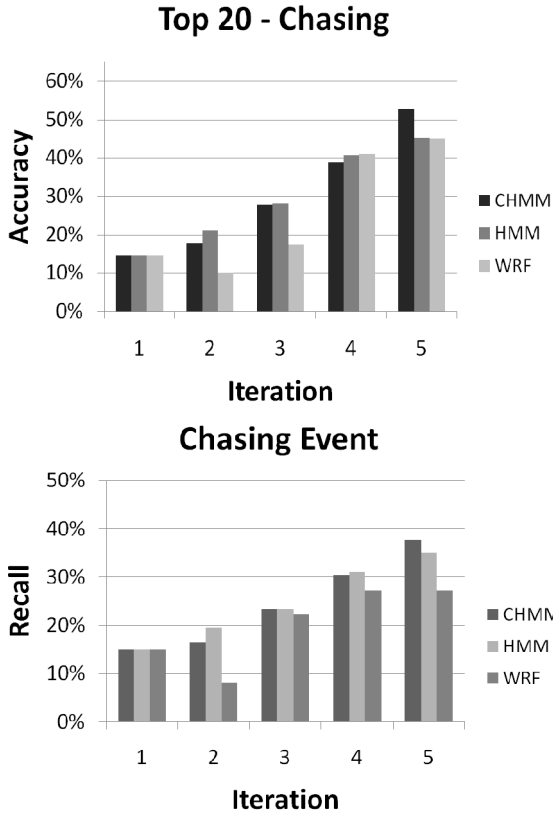

Figure 7. Chasing event retrieval accuracy and recall.

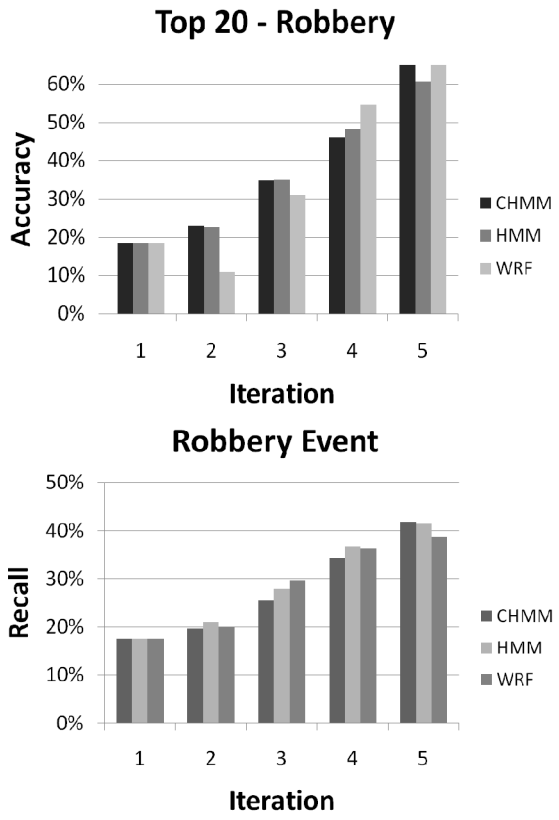

Figure 8. Robbery event retrieval accuracy and recall.

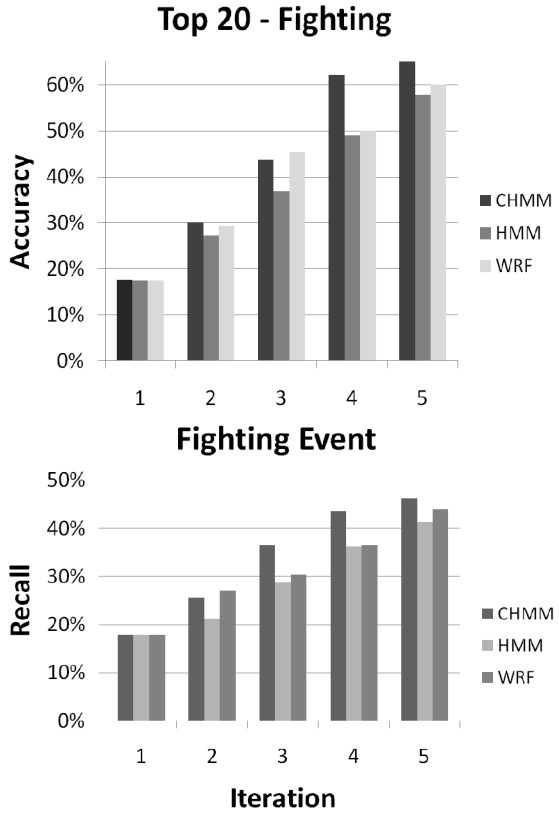

Figure 9. Fight event retrieval accuracy and recall.

of CHMM is about 9% higher than that of the HMM and 7% higher than that of the WRF.

2) Subtype Robbery Event Retrieval: We further compare the retrieval accuracy on the three types of robbery events. Similarly, the accuracy and recall values in the experimental results are calculated based on the 18 queries for Type I, 13 queries for Type II, and 13 queries for Type III, respectively.

The comparison results are illustrated in Figs. 10, 11, 12. For Type I robbery event, the accuracy values of CHMM, HMM, and WRF are 25.83%, 23.06%, and 0.00%, respectively at the 5th iteration. The corresponding recall values are 27.47% 25.00%, and 0.00%, respectively. For Type II robbery event, the accuracy values of CHMM, HMM, and WRF are 24.62%, 24.62%, and 0.00%, respectively at the 5th iteration. The corresponding recall values are 34.32%, 36.69%, and 0.00%, respectively. For Type III robbery event, the accuracy values of CHMM, HMM, and WRF are 56.54%, 46.15%, and 55.00%, respectively. The corresponding recall values are 76.33%, 66.86%, and 81.66%, respectively.

The accuracy results show that the CHMM performs better than that of the HMM and WRF in retrieving Type I and Type III robbery events. In retrieving Type II robbery event, CHMM and HMM perform almost equally well. The accuracy values of WRF are zeros in retrieving Type I and Type II robbery events due to no relevant CAI is returned at the second round of retrieval, and therefore,

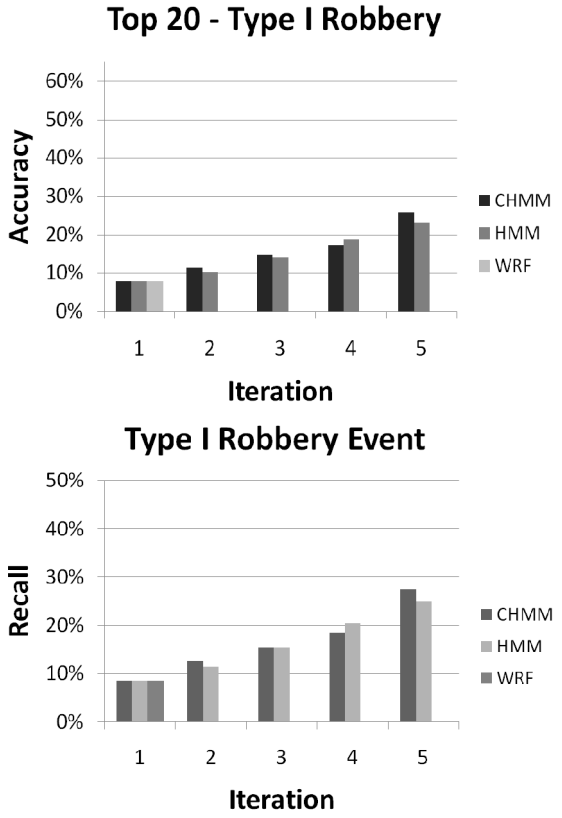

Figure 10. Accuracy and recall of Type I robbery event.

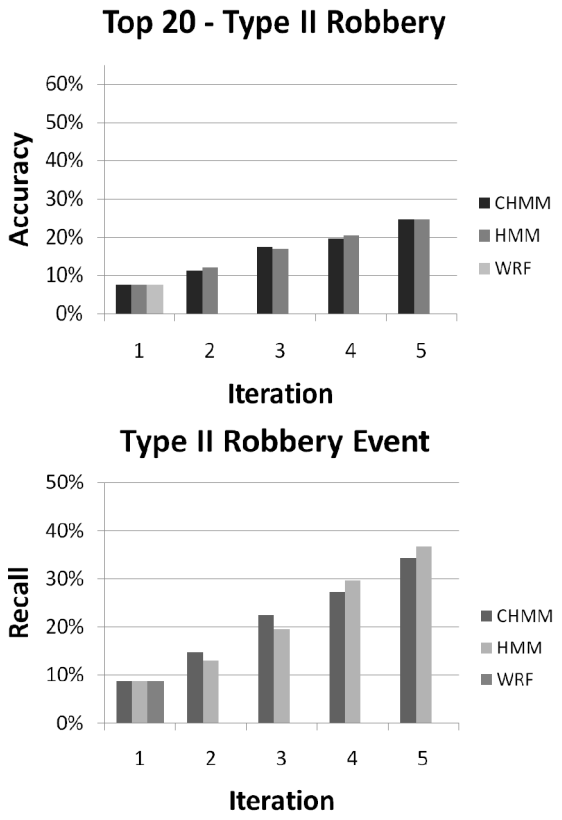

Figure 11. Accuracy and recall of Type II Robbery event.

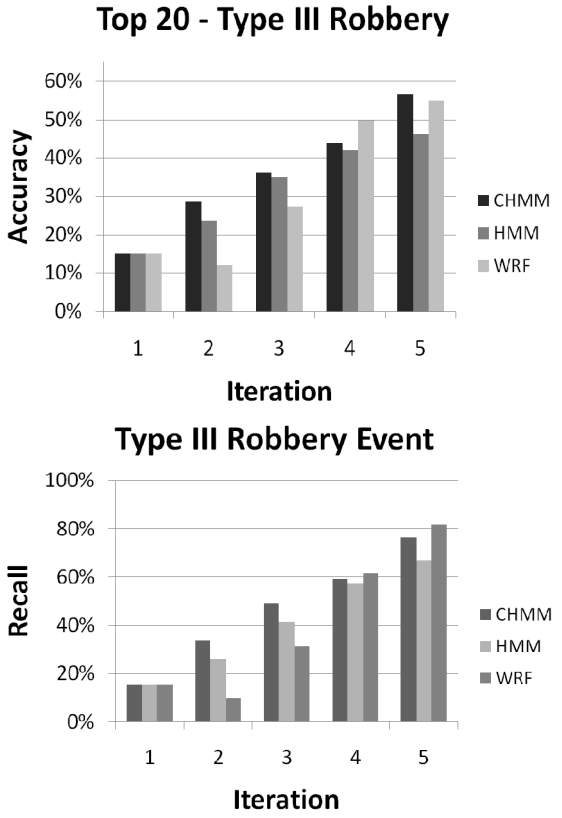

Figure 12. Accuracy and recall of Type III robbery event.

no users' feedback can be collected for updating the weighting scheme. As a consequence, there is no change in the subsequent retrieval results. In retrieving Type I robbery event, the recall value of the proposed method is better than that of the HMM and WRF. However, in retrieving Type II and III robbery events, the recall values of CHMM are lower than that of the HMM and WRF, respectively. This might be due to the fact that each subtype contains only a small number (less than 20) of subtype robbery events. The experimental results suggest that the proposed CHMM based method better models the spatio-temporal interaction between the robber(s) and the victim during the course of robbery.

In addition, similar performance charts can be observed from the experimental results in retrieving Type I and II robbery events. It is worth noting that Type I and II robbery events are similar to each other in terms of their macro and micro interactions. The only difference between these two event types is the number of objects involved. As a matter of fact, both Type I and Type II robbery are essentially the same since Type II robbery is actually Type I robbery plus an extra participant (robber). This is evidenced in our experimental results where the false positives in the retrieval of one subcategory events largely come from the other subcategory. Since the proposed learning method models events based on the pair-wise spatio-temporal relationships between two objects involved in the event, it may not be able to tell the difference between Type I and II robbery events. A possible solution is to use a set of CHMM models instead

of one when there are more than two objects being involved in certain type of event. The challenging issues would include how to determine the proper number of CHMM models in the set given a particular type of event, without any prior knowledge of the number of participants as well as the event structure. Another related and non-trivial issue is the extra computational cost associated with training and testing if a group of CHMMs are used. For an intelligent security surveillance system, the bottom line is to locate those events of interests and roughly classify them into big categories, though we admit that further studies on finer modeling of multi-object events would be worth pursuing as it could make a general impact to the related fields.

Fig. 13 demonstrates the overall retrieval accuracies averaged on all types of events across iterations within different scopes (top 5, 10, 15, and 20) for all three methods. In particular, five iterations of the average retrieval accuracies are shown within each scope. The experimental results show that the proposed CHMM framework has the highest average accuracy among all the three methods used for comparison. In addition, it is worth noting that almost all the three methods demonstrate an increase in their average accuracies across iterations after the initial query. This indicates that the relevance feed and multiple-instance learning can gradually promote relevant events and in the meanwhile demote irrelevant events.

By comparing the experimental results in this paper with the published results in our previous work [23], a worse result can be observed for fighting events probably due to the following reasons. First, the dataset used in the MIPR paper is much smaller than the one used in this paper. In addition, the video corpus used in the MIPR paper contains mainly fighting and normal events, which are relatively simple. However, the video corpus used in this paper contains various types of abnormal events, including fighting, chasing, three different types of robbery events, and normal events, where a much higher level of noise can be expected. Second, the learning and retrieval frameworks are quite different in these two papers. The system proposed in our MIPR paper uses CHMM to model object interactions directly based on sequence pairs where complete label information is available, i.e., the user provides feedback for individual SPs. In contrast, in this paper, the proposed system not only uses CHMM for event modeling, but also adds two levels of Multiple Instance Learning which enables the ability to perform query by example (QBE) at a higher indexing level (CAIs) for video event retrieval. Due to the added uncertainty and the less complete label information in a two-level MIL framework, the performance of the system proposed in this paper is expected to be worse than that with complete instance label information. For the above reasons, it is not proper and also not fair to directly compare the experimental results of the two frameworks.

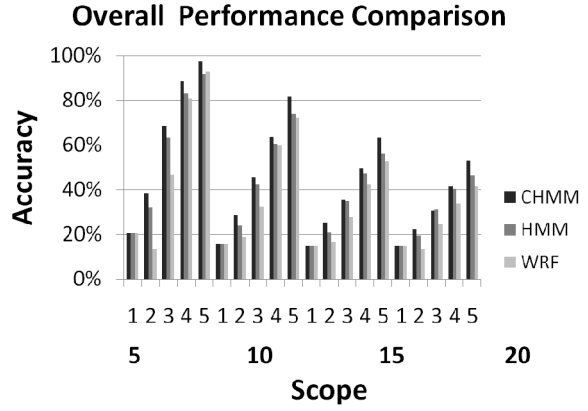

Figure 13. Overall retrieval accuracies across iterations.

IV. Conclusions

In this paper, we proposed a human-centered semantic video retrieval framework for searching various events in the surveillance video database. It is worth mentioning that the proposed framework not only takes advantage of coupled hidden Markov model in modeling various kinds of human interactions, but also tightly integrates the coupled hidden Markov models with relevance feedback and multiple-instance learning to enable the incorporation of the user's subjective perceptions of the retrieval result. To our best knowledge, our work is among the first effort to tackle the surveillance video retrieval problem this way.

In addition, by selecting proper features, this platform can analyze both macro and micro interactions in the video, which can be used to differentiate events that agree in macro interactions, but differ from each other in terms of micro interactions. The effectiveness of the proposed system is demonstrated by our experimental results.

In our future work, more general event models will be constructed and tested with the proposed platform. More videos containing other types of events will be collected to test the proposed framework.

Acknowledgment

The work of Chengcui Zhang was supported in part by the UAB ADVANCE program and NSF DBI-0649894.

References

[1] G. Lavee, L. Khan, and B. Thuraisingham, "A framework for a video analysis tool for suspicious event detection," *Multimedia Tools and Applications*, vol. 35, issue 1, pp. 109-123, 2007.

[2] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, "Semantic-based surveillance video retrieval," *IEEE Trans. Image Processing*, vol. 16, no. 4, pp. 1168-1181, Apr. 2007.

[3] I. Ersoy, F. Bunyak, and S.R. Subramanya, "A framework for trajectory based visual event retrieval," *in Proceedings of the International Conference on Information Technology: Coding and Computing* (ITCC 2004), Las Vegas, NV, Apr. 2004, pp. 23-27.

[4] A.F., Bobick, A.P. Pentland, and T. Poggio, "VSAM at the MIT Media Laboratory and CBCL: Learning and understanding action in video imagery," *in Proceedings of the DARPA Image Understanding Workshop* (DARPA 1997), New Orleans, LA, May. 1997, pp. 25-30.

[5] N. Brewer, N. Liu, O.D. Vel, and T. Caelli, "Using coupled hidden Markov models to model suspect interactions in digital forensic analysis," *in Proceedings of the International Workshop on Integrating AI and Data Mining* (IADM 2006), HongKong, China, Dec. 2006, pp. 58-64.

[6] V.M. Kettnaker, "Time-dependent HMMs for visual intrusion detection," *in Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop* (CVPRW 2003), Madison, Wisconsin, Jun. 2003, pp. 34.

[7] M. Petkovic and W. Jonker, "Content-based video retrieval by integrating spatio-temporal and stochastic recognition of events," *in Proceedings of the IEEE International Workshop on Detection and Recognition of Events in Video*, Vancouver, Canada, Jul. 2001, pp. 75-82.

[8] N.M. Robertson and I.D. Reid, "Behavior understanding in video: A combined method," *in Proceedings of the 10th IEEE International Conference on Computer Vision* (ICCV 2005), Beijing, China, Oct. 2005, pp. 808-815.

[9] A. Naftel and S. Khalid, "Classifying spatiotemporal object trajectories using unsupervised learning in the coefficient feature space," *Multimedia Systems*, vol. 12, no. 1, pp. 227-238, 2006.

[10] X. Chen, C. Zhang, and W.-B. Chen, "A multiple instance learning framework for incident retrieval in transportation surveillance video databases," *IEEE 23rd International Conference on Data Engineering Workshop*, Istanbul, Turkey, April 2007, pp. 75-84.

[11] J. Meessen, X. Desurmont, J.-F. Delaigle, C. De Vleeschouwer, and B. Macq, "Progressive learning for interactive surveillance scenes retrieval," *in IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, June 2007, pp. 1-8.

[12] S.-C. Chen, M.-L. Shyu, S. Peeta, and C. Zhang, "Learning-based spatio-temporal vehicle tracking and indexing for transportation multimedia database system," *IEEE Trans. on Intelligent Transportation Systems*, vol. 4, no. 3, pp. 154-167, 2003.

[13] Y. K. Jung, K. W. Lee, and Y. S. Ho, "Content-based event retrieval using semantic scene interpretation for automated traffic surveillance," *IEEE Trans. Intell. Transport. Syst.*, vol. 2, no. 3, pp. 151–163, 2001.

[14] S. Kamijo, Y. Matsushita, K. Ikeuchi, and M. Sakauchi, "Traffic monitoring and accident detection at intersections," *IEEE Trans. Intell. Transport. Syst.*, vol. 1, no. 2, pp. 108–118, 2000.

[15] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, "A system for learning statistical motion patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1450–1464, 2006.

[16] L. Chen and M. T. Özsu, "Modeling of Video Objects in a Video Database," *in Proceedings of the IEEE International Conference on Multimedia*, Lausanne, Switzerland. 2002.

[17] M. Petkovic and W. Jonker, "Content-based video retrieval by intergrating spatio-temporal and stochastic recognition of events," *in Proceedings of the IEEE International Workshop on Detection and Recognition of Events in Video*, Vancouver, Canada, 2001, pp.75-82.

[18] N. M. Robertson and I. D. Reid, "Behaviour understanding in video: a combined method," *in Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, Oct. 2005, pp. 808-815.

[19] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 8, pp. 873–879, 2001.

[20] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, Cambridge University Press, 2005.

[21] N.K. Kanhere, S.J. Pundlik, and S.T. Birchfield, "Vehicle segmentation and tracking from a low-angle off-axis camera," *in Proceedings of the 2005 IEEE Computer Society Interational Conference on Computer Vision and Pattern Recognition* (CVPR 2005), San Diego, CA, Jun 20-25. 2005, pp. 1152-1157.

[22] M. Brand, "Coupled hidden Markov models for modeling interacting processes," MIT Media Lab Perceptual Computing/Learning and Common Sense Technical Report 405 (Revised), June 1997.

[23] X. Chen, and C. Zhang, "Semantic event retrieval from surveillance video databases," *in Proceedings of the 2008 10th IEEE International Symposium on Multimedia*, San Francisco, CA, Dec. 2008, pp. 625-630.

Chengcui Zhang is an Assistant Professor of Computer and Information Sciences at The University of Alabama at Birmingham (UAB). She received her Ph.D. from the School of Computer Science at Florida International University, Miami, FL, USA in 2004. Her research interests include multimedia databases, multimedia data mining, image and video database retrieval, bioinformatics, and GIS data filtering.

Wei-Bang Chen is a Ph.D. candidate in the Department of Computer and Information Sciences at The University of Alabama at Birmingham. He received a Master's degree in Genetics from National Yang-Ming University in Taipei, Taiwan and a Master's degree in Computer Sciences from UAB. His main research area is bioinformatics. His current research involves microarray image and data analysis, biological sequence clustering, and biomedical video and image mining.

Xin Chen received her Ph.D. degree in Computer and Information Sciences from The University of Alabama at Birmingham, USA, in 2008. Her research interests include Content-based Image Retrieval, multimedia data mining, and spatiotemporal data mining.

Lin Yang is a Ph.D. student in the Computer and Information Sciences Department at The University of Alabama of Birmingham. His research is on mining useful information from community-contributed multimedia data using data mining, machine learning and computer vision techniques. He received the BS degree in computer science from Fudan University in 2006.

John K. Johnstone is an Associate Professor of Computer and Information Sciences at UAB. He received his B.Sc. in Mathematics from the University of Saskatchewan, and his M.S. and Ph.D. in Computer Science from Cornell University. He has also been on the faculty at Johns Hopkins University. His primary research interest is shape modeling, with recent interest in the interface of graphics and vision. His research has been supported by several NSF grants.